Int J Digit Libr (2018) 19:127–138
https://doi.org/10.1007/s00799-017-0225-7

The context of multiple in-text references and their signification

Marc Bertin[1] · **Iana Atanassova**[2]

Received: 18 October 2016 / Revised: 8 July 2017 / Accepted: 11 July 2017 / Published online: 21 July 2017

Abstract In this paper, we consider sentences that contain multiple in-text references (MIR) and their position in the rhetorical structure of articles. We carry out the analysis of MIR in a large-scale dataset of about 80,000 research articles published by the Public Library of Science in 7 journals. We analyze two major characteristics of MIR: their positions in the IMRaD structure of articles and the number of in-text references that make up a MIR in the different journals. We show that MIR are rather frequent in all sections of the rhetorical structure. In the Introduction section, sentences containing MIR account for more than half of the sentences with references. We examine the syntactic patterns that are most used in the contexts of both multiple and single in-text references and show that they are composed, for the most part, of noun groups. We point out the specificity of the Methods section in this respect.

Keywords Multiple in-text references · Bibliometrics · Citation analysis · In-text references · Content citation analysis · IMRaD structure · POS tagging

✉ Marc Bertin
bertin.marc@gmail.com

Iana Atanassova
iana.atanassova@univ-fcomte.fr

[1] Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal (UQAM), Montréal, Canada

[2] Centre CRIT-Tesnière, University of Bourgogne Franche-Comté (UBFC), Besançon, France

1 Introduction

Scientific articles are highly structured texts and exhibit many regularities related to their rhetorical structure. Considering the frequency of in-text references, this is highly dependent on the rhetorical structure. For example, Bertin et al. [3] study the differences between the four sections of the IMRaD (Introduction, Methods, Results and Discussion) structure in terms of the density of citations.

Our approach allows the identification of multiple in-text references in texts. In general, the presence of more than one in-text references in the same sentence gives us information about a relative proximity between the works that are cited. Previous studies on in-text references take into account word windows or sentences to study citation contexts. However, some recent works show the importance and the difficulty in identifying citation blocks that are spans of citations that may encompass one or more sentences [15]. Another related question, the recurrence of in-text references or re-citations, has been the object of several studies [1,14,26].

We focus on text spans in articles that contain more than one in-text references that appear very close to each other. We consider sentences as a basic textual unit, and we examine sentences containing more than one in-text reference. We call this phenomenon *multiple in-text references (MIR)*. For example, the following sentence contains a MIR composed of 4 references, where 3 of the references appear in the range *"[74–76]"*:

> *Indeed, it has long been proposed on thermodynamic grounds that transcription factors would bind at low, nonfunctional levels throughout the genome either via sequence-independent [74–76] or sequence-specific DNA binding [32].*[1]

[1] PLOS Biology, 2008, DOI: 10.1371/journal.pbio.0060027.

In general, MIR can appear in several different groups in the same sentence, that we call *aggregates of in-text references*. The aggregates are either enumerations of in-text references or ranges as in the example above. A sentence containing MIR can contain one or more aggregates separated by text.

In this paper, we present the results on the behavior of MIR, their positions in the IMRaD structure, and their contexts, that we obtain by processing a large-scale corpus of about 80,000 research articles. The key idea involves identifying the number of in-text references at the level of the sentence. We analyze the relations between three major characteristics of MIR: their positions in the IMRaD structure, the number of in-text references that make up a MIR and the syntactic patterns that appear in the contexts around MIR. The motivation for this approach is the search for mechanisms of category assignment to in-text references. According to Cronin and Wouter [4,25], this last point is essential, from the semiotic point of view, to the foundation of a theory of citation.

2 Related work

The phenomenon of multiple in-text references, i.e., references that appear in the same sentence, in scientific papers has not yet been studied in terms of their positions and contexts. In recent works on the rhetorical structure of articles [3,9], the studies are based on the presence of in-text references in sentences but their number in a single sentence is not taken into consideration. Several works exist on a related problem which is the proximity of co-citations in texts [13]. Liu and Chen [17,18] propose a four-level co-citation proximity scheme for the levels of article, section, paragraph and sentence. Ding [10] proposes a content-based citation analysis approach based on the use of context at both the syntax and semantic levels. A related study by Day (see [7]) highlights the encoding of psychological frameworks of human behavior through citations. From the linguistic point of view, a method has been proposed to measure the difference in relative frequencies of specific socio-epistemic terms and phrases in dissertation abstracts [8], but to our knowledge this method has not yet been applied to the study of citation contexts. Other points of view to citation contexts are developed in the collection by Cronin and Sugimoto [6].

Cronin [5] states that "*the meaning of the citation*" is the essential element for citation theory. The semiotic approach of Pierce has been applied to the triad of citation by Wouters [25]. This approach takes into consideration the "*sign*" as a structural element of the citation triad which is a fundamental element of the theory of citation that remains to be developed. Wouter [25, p. 71] suggests, in the article *Semiotics and citation*, an essential argument on the nature of citation and reference:

> By analyzing references and citations as different signs, they were essentially positioned as different objects. Their relation of descent: the citation emerges in an act of semiosis (the creation of a novel sign) form the reference.

In this article, the essential point is the fact that we propose a method to define more precisely a typology of the contexts of in-text references to help us understand the semiotic nature of citations. If the nature of this study is applicative, it is important to note that it is related to the foundations of a theory of citation. An essential point of this study is the fact that we formulate the hypothesis that the proximity of in-text references is important to the attribution of the constitutive semantic values of the relation. We suppose that the aggregates of in-text references, that are presented in the form of enumeration of references, bear, by their position, the same attributes. They form a composite object that has the same properties as a simple reference but points to several citations. The fundamental question from a metrical and symbolic point of view, but also in terms of networks, remains to be studied. In this paper, we focus on the identification of syntactic patterns around in-text references.

3 Method

To address the problem of the identification of MIR and their positions in the rhetorical structure of articles, we have processed a large-scale corpus of articles that follow the IMRaD structure. In fact, during the last decades, IMRaD has imposed itself as a standard rhetorical framework for scientific articles in the experimental sciences.

The objective of our study is to determine the locations where MIR are most likely to appear in the rhetorical structure of scientific articles. We also study the number of in-text references in the sentences.

3.1 Dataset

To perform this study, we have analyzed a dataset of seven peer-reviewed academic journals published in Open Access by the Public Library of Science.[2] Six of the journals are domain-specific (*PLOS Biology, PLOS Computational Biology, PLOS Genetics, PLOS Medicine, PLOS Neglected Tropical Diseases*) and the seventh is *PLOS ONE*, which is a general journal that covers all fields of science and social sciences. We have processed the entire dataset of about 80,000 research articles in full text published up to September 2013.

[2] PLOS, http://www.plos.org.

The dataset is in the XML JATS format, where the body of each article consists of sections and paragraphs that are identified as distinct XML elements. The in-text references are also, for the most part, marked up as XML elements and linked to the corresponding elements in the bibliography of the article.

3.2 Identification of the IMRaD structure

To identify the IMRaD structure of articles, we have analyzed all section titles and categorized the sections. All seven journals use similar publication models, where authors are explicitly encouraged to use the IMRaD structure. As a result, more than 97% of the research articles in the corpus contain the four main section types (Introduction, Methods, Results and Discussion), although not always in the same order. While the relative position of a section may have an influence on the number of references found in the section, e.g., the Discussion section that appears immediately after an Introduction section may have more references than the Discussion section that appears at the end of the article, the order in which the sections appear in an article has not been taken into consideration for this study. The detailed results of the analysis of the IMRaD structure for this corpus are presented by Bertin et al. [3].

3.3 Identification and processing of MIR

In order to identify sentences containing MIR, we need to perform the following steps:

1. Segment all paragraphs into sentences;
2. Identify all in-text references;
3. Consider the number of in-text references in each sentence.

The first step was done by analyzing the punctuation and capitalization of the text in order to identify sentence ends. Our corpus contains a total of 15,852,120 sentences, out of which 3,528,514 (around 22%) contain in-text references.

The second step may seem trivial given that the in-text references are present as elements in the XML tree of the article. In the case of MIR however, this is not always true. When in-text references are in a numeric form, reference ranges are often present in sentences containing MIR. For example, we can consider the following sentence in the corpus with XML markup:

> *A number of recent studies have used a modification of the picture viewing procedure by substituting pleasant pictures with photographs of loved, familiar faces* <xref ref-type="bibr" rid="pone.0041631-Bartels1">*[16]*</xref>–<xref ref-type="bibr" rid="pone.0041631-Xu2">*[24]*</xref>.[3]

The in-text references in this sentence "[16]–[24]" are identified as two *xref* elements that point to the corresponding bibliography items. In reality, the sentence contains 9 different citations: all the works from [16] to [24] are cited; and 7 of these citations are not present in the XML markup. In such cases, we call *implicit in-text references* those in-text references that are part of a range but that are not mentioned by their numbers in the sentence.

In order to identify correctly MIR and their number in sentences, it is important to detect in-text reference ranges and implicit references. To do this, we first examine the context of each *xref* element and identify all possible ranges. Then, we generate the list of implicit references and the links between these references and the bibliography items. A similar method for the processing of in-text reference ranges was used by Bertin et al. [3].

The occurrences of in-text reference ranges are rather numerous in our corpus. We found out that reference ranges are present in 19.19% of all sentences containing MIR and implicit in-text references account for 12.38% of the MIR.

3.4 Types of in-text references

A related problem is the identification of in-text references that do not appear as elements in the XML tree of the article. In fact, the PLOS corpus contains a small amount of such unidentified in-text references (less than 0.02%). However, this problem is important from methodological point of view because it is a major issue when working with other corpora that are not in XML JATS format.

The identification of in-text references in general is closely related to the problem of named entity recognizer (NER). The main idea is to label sequences of words in texts which represent the names of persons, company names or gene and protein names. From a technical point of view, Stanford NER uses CRF Classifier [12]. Figure 1 shows examples of representations of sentences containing MIR. These examples are generated using Brat,[4] which is a web-based tool for text annotation [24].

As shown in these examples, in-text references have various structures. They can be alphanumeric but also digital. In order to identify them, it is necessary to build new resources by machine learning to take into account the specifics of the different representations of in-text references. Typography and context play an important role when determining the borders of in-text references. In the example (a) in Fig. 1, we have a named entity (*"Borneo"/LOC*) which is not part of an in-text reference. Similarly, in the examples (b) and (c), there are numeric and alphanumeric tokens that appear at various positions in the sentences and that

[3] PLOS ONE, 2012, DOI:10.1371/journal.pone.0041631.

[4] http://brat.nlplab.org.

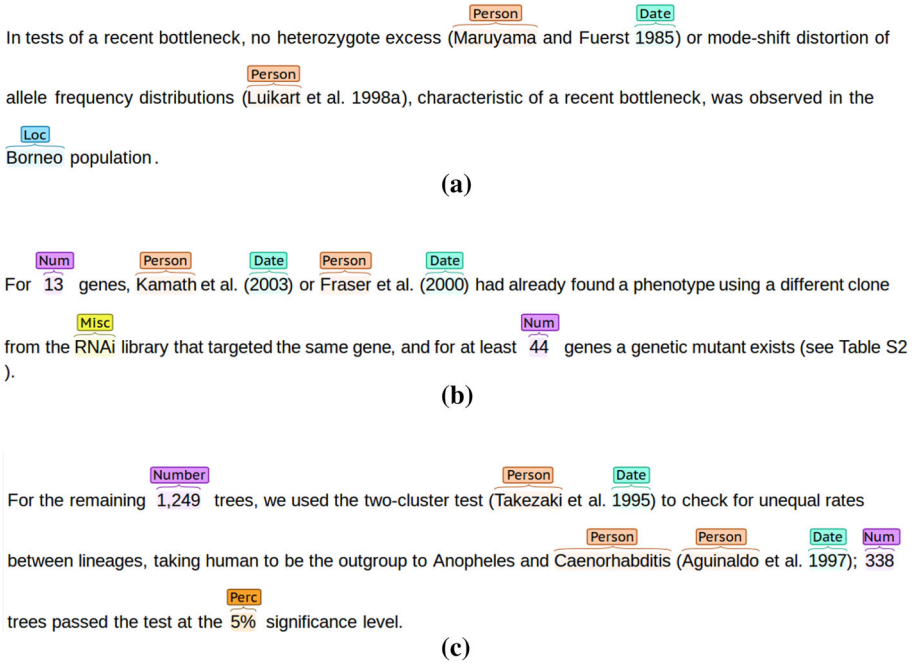

Fig. 1 Named entity recognition examples

could introduce noise in the in-text reference identification.

Figure 2 presents the different types of in-text references that can be found in articles. The types of references are ordered from the left to the right by their growing morphological complexity. The numeric references have the most simple form and are the easiest to identify. There exist several different forms of alphanumeric references that combine author names, dates, some other characters and specific punctuation. The morphological complexity is inversely proportional of the understanding of the reference out of context, i.e., independently of the bibliography provided at the end of the article. Numeric references have no significance out of context, because the cited work cannot be identified without the bibliography. Alphanumeric references can be understood, according to their form, by researchers in the same field of study or more broadly in the same discipline. Finally, we have considered the extreme case of in-text references that are part of the text as named entities and that do not necessarily appear in the bibliography, such as *"Einstein"* or *"Darwin."* Such references refer to works that are universally known to researchers.

A robust method for the identification of in-text references needs to take into consideration all the different customs in writing in-text references, and these can vary according to the discipline. We note that there exist international standards that provide guidelines for bibliographic references, namely ISO 690:2010, that are necessary to carry out the text processing on a large-scale corpus. In general, the morphology of MIR is more complex and presents more variations than single in-text references.

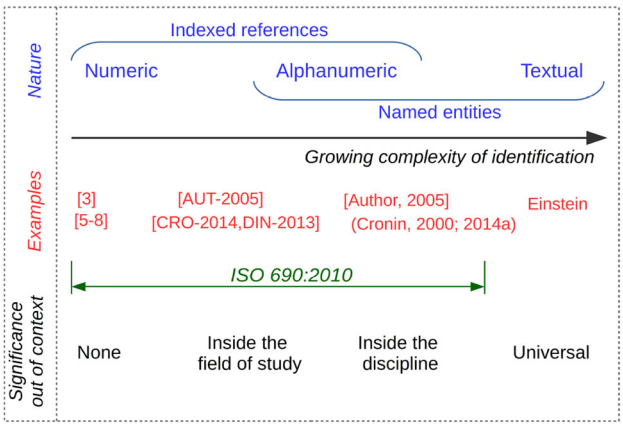

Fig. 2 Types of in-text references

3.5 Categorization and natural language processing approaches

For the natural language processing (NLP) of MIR context analysis, we use the Stanford CoreNLP, which is a suite of core NLP tools [19]. We focus on the part-of-speech (POS) tagger, the named entity recognizer (NER) and the coreference resolution system to show relevant information located in MIR contexts and to demonstrate the relevance of these problems for citation context analysis. According our protocol, we extract only those sentences in the articles that contain multiple in-text references and we use them to show the complexity of the relations between the in-text references. The most basic level of annotation is the part-of-speech level, as shown in Fig. 3.

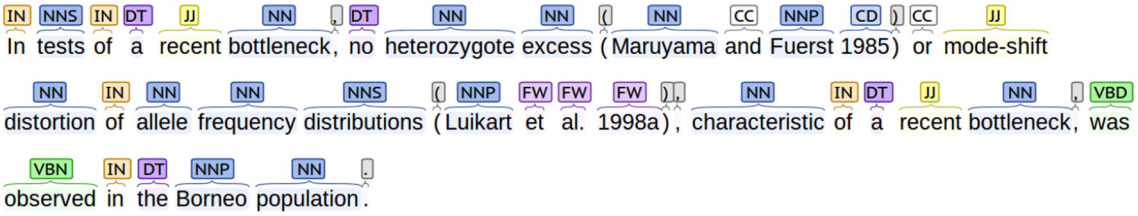

Fig. 3 Part-of-speech tags: example

Table 1 MIR in the IMRaD Structure

	I (%)	M (%)	R (%)	D (%)	Total (%)
Sentences with MIR	52.78	25.05	35.59	42.65	41.43
Sentences with a single reference	47.22	74.95	64.41	57.35	58.57

We note that the in-text references are considered as named entities by the POS tagger (see Fig. 3). To avoid possible errors, we have tagged all in-text references that were identified by the previous processing by *xref* tags so that they are not part of the tagger input. We obtained the sequences of POS tags that appear in the contexts of in-text references. This allowed us to examine the most frequent types of contexts in terms of their syntactic structures. The results are presented in the Sect. 4.3.

4 Results

We first observe the presence of MIR in the four section types of the IMRaD structure, and then, we examine the number of the MIR according to the different journals in the dataset.

4.1 Use of MIR in the IMRaD structure

Table 1 presents the percentage of sentences containing multiple in-text references (MIR) among all sentences with in-text references in the four section types of the IMRaD Structure (I-M-R-D). We observe that MIR are present for the most part in the Introduction section where more than half of the sentences with in-text references contain MIR (52.78%). This result is consistent with the observation that the Introduction section often includes a state of the art with a literature review in which MIR are most likely to appear. As for the Methods and Results sections, they have around 25 and 35% of MIR, respectively.

The results in Table 1 for the four different section types are not unexpected. In fact, the relative quantity of MIR in the sections follows the overall distribution of references in the IMRaD structure, shown in Bertin et al. [3], where the Methods section contains the smallest number of in-text references, followed by the Results section. The Introduction section, which is also the shortest one on average, contains the highest number of citations. This table shows also that MIR appear very often: in around 41% of all sentences containing in-text references. This means that in a scientific article, the largest number of in-text references appears in groups of several references situated closely in the textual space, i.e., in the same sentence.

Table 2 Percentage of sentences with MIR in the IMRaD structure

Number of in-text references in MIR	I (%)	M (%)	R (%)	D (%)
2	21.74	15.67	19.49	20.74
3	11.63	4.43	7.31	9.34
4	6.25	1.53	3.05	4.36
5	3.45	0.68	1.39	2.12
6	2.04	0.34	0.70	1.13
7	1.22	0.17	0.39	0.64
8	0.77	0.10	0.23	0.36
9	0.48	0.06	0.14	0.21
10	0.32	0.05	0.10	0.13
11	0.22	0.03	0.07	0.08
12	0.15	0.02	0.05	0.06
13	0.10	0.02	0.04	0.04
14	0.08	0.01	0.03	0.03
15	0.05	0.01	0.03	0.02

Table 2 presents the percentage of sentences with MIR of different sizes among all sentences containing citations in each of the four section types. Considering the MIR with 2 elements, we observe that there is little difference between the sections Introduction, Results and Discussion. Then, the differences between the sections increase with the number in in-text references. This means that, while MIR with 2 elements appear almost homogeneously in an article, the MIR with higher number of elements are more and more exclusively reserved to the Introduction section. This phenomenon again is explained by the presence of the state of the art in the Introduction with a very high concentration of in-text references.

Table 3 MIR in the 7 PLOS journals: percentage of articles that contain sentences with at least N references

Journal	$N = 1$ (%)	$N = 2$ (%)	$N = 3$ (%)	$N = 4$ (%)	$N = 5$ (%)	$N = 10$ (%)	$N = 15$ (%)	$N = 20$ (%)
PLOS Biology	100.00	100.00	99.37	95.50	84.38	18.98	3.71	1.03
PLOS Comp. Biology	100.00	100.00	99.49	95.51	85.98	25.31	6.76	1.91
PLOS Genetics	100.00	99.97	99.44	95.58	83.68	17.28	3.43	0.91
PLOS Medicine	100.00	100.00	98.27	88.88	75.92	22.25	10.80	6.91
PLOS N. Trop. Diseases	100.00	99.89	97.28	90.06	74.25	15.12	3.20	1.01
PLOS Pathogens	100.00	99.97	99.56	94.62	81.45	15.42	2.35	0.71
PLOS ONE	99.95	99.90	98.59	91.34	76.64	17.26	4.34	1.60

Table 4 MIR in the 7 PLOS journals: number of elements

Journal	Average number of elements in MIR	SD	Maximal number of elements in MIR
PLOS Biology	3.05	1.78	35
PLOS Computational Biology	3.22	2.09	64
PLOS Genetics	3.01	1.74	52
PLOS Medicine	3.32	3.81	190
PLOS Negl. Tropical Diseases	3.05	1.86	46
PLOS Pathogens	2.98	1.67	46
PLOS ONE	3.13	2.14	288

Table 5 Examples of sentences with high number of in-text references

Journal	Section	Sentence
PLOS Medicine	R	The systematic review identified 188 studies that provided prevalence estimates [18,29,36–223]
PLOS Medicine	M	Data for calculation of the number of snakebite envenomings were obtained for 46 countries [9–60] while data for calculation of the number of deaths due to snakebite were obtained for 22 countries [9–11,18, 31,47,49,51,56,58–72] by this process
PLOS ONE	R	As a result, 221 unique genes and 4 protein complexes (DNA-PK, HSP70, MRN(95), RAS) were identified from around 200 papers that studied radiation response-related biomarkers [4], [14]–[185]
PLOS ONE	M	Details of each study [11]–[113] were entered into a database by one investigator with a 100% re-check

4.2 MIR in the PLOS journals

Table 3 presents the percentage of articles in each journal that contain sentences with at least N in-text references. We observe some differences between the journals in the use of very large MIR (with number of elements 10, 15 and above). In fact, the journal PLOS Medicine stands out because it uses MIR with relatively high number of elements: as much as 22% of its articles contain sentences with 10 or more in-text references, and 11% with 15 or more in-text references. In PLOS Computational Biology, the use of MIR with at least 10 elements is quite common (around 25% of the articles) but only around 7% of the articles use MIR with 15 or more references.

Table 4 presents the average and the maximal number of elements in MIR observed in the 7 journals. PLOS Medicine has the highest average number of elements in MIR. In fact, articles in this journal tend to be short, but with a very high number of references and many of them appearing in the same sentence. PLOS ONE and PLOS Medicine have very high maximal number of elements in MIR. However, as shown in Table 3, MIR with a high number of elements in PLOS ONE tend to be less frequent than those in PLOS Medicine.

Some examples of sentences containing MIR with very high number of In-text References are presented in Table 5. In fact, these examples are extracted from articles in the medical domain that are of a specific type: systematic reviews. This kind of articles has for objective to sum up the best available

Table 6 Most frequent tag sequences per section in the context of multiple in-text references (MIR)

Section	Patterns around in-text references	Frequency (%)
Introduction	NN IN DT JJ NN < xref − range >	0.23
	NN IN JJ JJ NNS < xref − range >	0.18
	DT NN IN JJ NNS < xref − range >	0.16
	JJ NN IN JJ NNS < xref − range >	0.13
	NN IN JJ NN NNS < xref − range >	0.13
	JJ NN IN NN NNS < xref − range >	0.12
	NN NN LRB NN RRB < xref − range >	0.12
	DT JJ NN IN NN < xref − range >	0.11
	JJ NNS IN DT NN < xref − range >	0.11
	JJ NNS IN JJ NNS < xref − enum >	0.11
Methods	VBD VBN IN RB VBN < xref − enum >	0.40
	VBD VBN IN VBN RB < xref − enum >	0.20
	NN LRB NN RRB < xref > , NN LRB NN RRB	0.10
	NN VBD VBN IN VBN < xref − enum >	0.10
	NNS VBP VBN VBN RB < xref − enum >	0.10
	IN DT JJ JJ NN < xref − enum >	0.09
	JJ NN IN JJ NNS < xref − enum >	0.09
	LRB VB, FW, < xref − enum > RRB	0.09
	NN, IN RB VBN < xref − enum >	0.09
	NN IN DT JJ NN < xref − enum >	0.09
Results	NN < xref − enum > , < xref − range > , < xref >	0.19
	NN IN DT NN NN < xref − enum >	0.17
	NN IN DT JJ NN < xref − enum >	0.17
	NNS < xref − enum > , < xref − range > , < xref >	0.16
	IN DT NN IN NN < xref − enum >	0.13
	IN JJ NNS < xref − range > , < xref >	0.13
	NN IN JJ NN NNS < xref − enum >	0.13
	NN LRB NN NN RRB < xref − enum >	0.13
	JJ NN IN JJ NNS < xref − enum >	0.12
	IN DT JJ NN NN < xref − enum >	0.10
Discussion	IN JJ NNS < xref > , < xref − range >	0.22
	NN IN DT JJ NN < xref − enum >	0.22
	NN IN DT NN NN < xref − enum >	0.20
	JJ NN IN JJ NNS < xref − enum >	0.18
	JJ NN NNS < xref > , < xref − range >	0.18
	NN IN JJ NN NNS < xref − enum >	0.18
	IN NN NN < xref > , < xref − range >	0.14
	IN DT JJ NN NN < xref − enum >	0.13
	DT NN IN NN NN < xref − enum >	0.13
	JJ JJ NNS < xref > , < xref − range >	0.13

research on a specific topic by collecting and synthesizing the results of other studies that fit pre-specified eligibility criteria. For this reason, we find in these articles sentences that cite a large number of other works. Moreover, as shown in Table 5, these sentences are not necessarily in the Introduction section but appear quite often in the Results and Methods sections.

4.3 The contexts of in-text references

In this section, we examine the contexts of MIR in respect of the most frequent syntactic structures in which they appear. For the sake of comparison, we also look at the contexts of single in-text references (SIR).

Table 7 Most frequent tag sequences per section in the context of single in-text references (SIR)

Section	Patterns around SIR	Frequency (%)
Introduction	NN IN DT JJ NN < xref >	0.78
	NN IN DT NN NN < xref >	0.64
	DT NN IN JJ NNS < xref >	0.54
	JJ NN IN NN NN < xref >	0.48
	NN IN JJ NN NNS < xref >	0.46
	JJ NN IN JJ NNS < xref >	0.46
	DT NN IN NN NN < xref >	0.44
	IN DT JJ NN NN < xref >	0.44
	IN DT NN IN NN < xref >	0.44
	JJ NN IN DT NN < xref >	0.44
Methods	VBD VBN IN RB VBN < xref >	1.66
	VBN IN NNP FW FW < xref >	1.28
	VBD VBN IN VBN RB < xref >	1.00
	NN NN IN RB VBN < xref >	0.52
	NN VBD VBN IN VBN < xref >	0.46
	NN IN DT NN NN < xref >	0.44
	NN NN LRB NN RRB < xref >	0.44
	NN IN DT JJ NN < xref >	0.38
	NNS VBD VBN IN VBN < xref >	0.38
	VBN VBG DT NN NN < xref >	0.38
Results	NN IN DT NN NN < xref >	0.86
	NN IN DT JJ NN < xref >	0.60
	IN DT NN NN NN < xref >	0.56
	NN LRB NN NN RRB < xref >	0.40
	IN DT JJ NN NN < xref >	0.38
	NNS IN DT JJ NN < xref >	0.38
	NN IN JJ NN NNS < xref >	0.34
	VBN IN DT NN NN < xref >	0.34
	JJ NN IN JJ NN < xref >	0.32
	JJ NN IN NN NN < xref >	0.32
Discussion	NN IN DT JJ NN < xref >	0.90
	NN IN DT NN NN < xref >	0.60
	JJ NN IN JJ NNS < xref >	0.54
	JJ NN IN DT NN < xref >	0.50
	IN DT JJ NN NN < xref >	0.46
	JJ NN IN NN NN < xref >	0.46
	VBN IN DT JJ NN < xref >	0.44
	JJ NN IN JJ NN < xref >	0.42
	DT NN IN JJ NNS < xref >	0.36
	NN NN IN JJ NNS < xref >	0.36

After carrying out the POS tagging of the sentences containing MIR and SIR, we have considered the sequences of POS tags that appear in their contexts. Table 6 shows the top 10 most frequent syntactic patterns (or tag sequences) for the four different section types of the IMRaD structure. The third column gives the frequency of each pattern as a percentage of all patterns occurring in the section.

The tags used in this table are defined as follows:[5]

– < xref >: a single in-text reference, e.g., "[4]";

[5] The full definition of the tagset used by the Stanford POS tagger can be found at https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

Table 8 Most frequent tag sequences in the context of multiple in-text references per number of in-text references (MIR)

Number of in-text references in sentence (N)	Patterns around in-text references	Frequency (%)
$N = 2$	NN IN DT JJ NN < xref − enum >	0.55
	NN IN DT NN NN < xref − enum >	0.53
	VBD VBN IN RB VBN < xref − enum >	0.43
	JJ NN IN NN NN < xref − enum >	0.40
	DT NN IN JJ NNS < xref − enum >	0.38
	NN IN JJ NN NNS < xref − enum >	0.33
	IN DT JJ NN NN < xref − enum >	0.25
	IN DT NN IN NN < xref − enum >	0.25
	NNS IN JJ NN NNS < xref − enum >	0.25
	IN NN IN JJ NNS < xref − enum >	0.25
$3 \leq N \leq 4$	JJ NN IN JJ NNS < xref − enum >	0.24
	NN IN DT JJ NN < xref − enum >	0.23
	NN IN DT NN NN < xref − enum >	0.18
	NN IN JJ NN NNS < xref − enum >	0.16
	NN IN DT JJ NN < xref − range >	0.16
	VBN IN DT JJ NN < xref − range >	0.15
	NN IN JJ JJ NNS < xref − range >	0.11
	JJ NN IN JJ NNS < xref − range >	0.11
	NNS IN DT JJ NN < xref − enum >	0.11
	NN NNS IN JJ NNS < xref − enum >	0.11
$5 \leq N \leq 8$	NN IN DT JJ NN < xref − range >	0.13
	DT NN IN JJ NNS < xref − range >	0.13
	NNS < xref − enum > , < xref − range > , < xref >	0.13
	NN NNS IN JJ NNS < xref − range >	0.09
	NN IN JJ JJ NNS < xref − range >	0.08
	JJ NN IN JJ NNS < xref − range >	0.08
	NN NN CC NN NN < xref − range >	0.08
	NNS IN DT JJ NN < xref − range >	0.08
	DT NN IN JJ NN < xref − range >	0.08
	NN , NN CC NN < xref − enum >	0.08
$N \geq 9$	LRB VB , FW , < xref − enum > RRB	0.15
	DT NN IN JJ NNS < xref − range >	0.10
	NN < xref − enum > , < xref − range > , < xref >	0.10
	NN IN JJ NN NNS < xref − range >	0.10
	IN JJ NNS < xref > , < xref − range >	0.10
	NN IN JJ JJ NNS < xref − range >	0.09
	NN , NN CC NN < xref − range >	0.09
	NNS IN JJ NN NNS < xref − range >	0.09
	IN JJ CC JJ NNS < xref − range >	0.09
	NN IN DT JJ NN < xref − range >	0.07

- < xref − enum >: an aggregate formed by an enumeration of in-text references, e.g., "[2], [5]";
- < xref − range >: an aggregate formed by a range of in-text references, e.g., *"[8–10]"*;
- NN, NNS: noun (singular, mass or plural);
- JJ: adjective;
- IN: preposition or subordinating conjunction;
- DT: determiner;

– VBD, VBN: verb, past tense or past participle;
– VBP: verb, singular present;
– RB: adverb.

The contexts in the table were extracted from all sentences containing MIR and by taking into consideration a window a 5 tokens (words and punctuation signs) around in-text references. In almost all of the contexts in Table 6, the in-text references are at final position, which means that they appear most frequently at the end of the sentences. We observe that this is the case for most of the contexts of all four section types.

Considering the parts of speech that appear in these contexts, they belong mostly to noun groups for the sections Introduction, Results and Discussion. In these three sections, in-text references are for the large part preceded by sequences of nouns (NN, NNS), adjectives (JJ) and determiners (DT). This is the case when the references are placed directly after a noun group that can designate a result, a theory, the name of a method or a tool. The Methods section displays different kinds of contexts: in-text references in the Methods section tend to be surrounded by verb forms (VBD, VBN) and adverbs (RB). This means that, from a syntactic point of view, the in-text references in this section appear as part of the verb group in the sentence. This result should be read in the light of other studies on the contexts of in-text references. For example, Bertin et al. [2] propose a lexical analysis of citation contexts in the IMRaD structure and point out that the Methods section differs from the other sections in terms of the verbs that are employed around in-text references . The most frequent verbs in the Methods section are *"use," "perform," "follow," "obtain," "generate,"* and they appear rarely in the other three sections. The results in Table 6 are explained by the fact that these verbs, when in the context of in-text references, are most often used in the passive voice (e.g., *"was/VBD used/VBN"*) and closely followed by the in-text reference.

The results in Table 6 should be compared to the patterns of sentences containing single in-text references (SIR). Table 7 gives the tag sequences per section in the context of SIR and their frequencies. The higher frequencies for the top ten patterns around SIR mean that the contexts in which SIR appear are less diverse compared to the contexts of MIR in terms of syntactic structure. We observe the same specificity of the Methods section: the contexts of SIR in this section contain most often verb forms (VBD, VBN, VBG) and adverbs (RB). In particular, the most frequent pattern in the Methods section is *VBD VBN IN RB VBN* < xref > and it accounts for 1.66% of all occurrences of SIR in this section. This same pattern is also the most frequent in the case of MIR (with frequency 0.40%). Examples of such sentence are:

Table 9 Most frequent tag sequences in the context of single in-text references (SIR)

Patterns around SIR	Frequency (%)
NN IN DT JJ NN < xref >	0.67
NN IN DT NN NN < xref >	0.64
VBN IN NNP FW FW < xref >	0.48
VBD VBN IN RB VBN < xref >	0.42
IN DT JJ NN NN < xref >	0.40
VBN IN DT JJ NN < xref >	0.36
JJ NN IN DT NN < xref >	0.36
JJ NN IN NN NN < xref >	0.34
IN DT NN NN NN < xref >	0.33
JJ NN IN JJ NNS < xref >	0.32

Chromosomes counting ***was/VBD performed/VBN as/IN previously/RB described/VBN*** *[22].*[6]
Analysis of binding antibodies against whole SARS-CoV lysates or N protein, and S glycoprotein-specific neutralizing antibody (Nab) ***were/VBD conducted/VBN as/IN previously/RB reported/VBN*** *[11].*[7]

Further, we examined how the number of in-text references in the sentences impacts the syntactic structure of the contexts. Table 8 presents the top 10 most frequent syntactic patterns depending on the number of in-text references in the sentence. The frequency of each pattern is given as a percentage of all patterns occurring around the same number of in-text references. For the sake of comparison, Table 9 gives the tag sequences in the context of SIR for all sections.

We observe that the patterns containing verb groups appear frequently in sentences with 1, 2, 3 or 4 in-text references, but are less frequent for $N \geq 5$. All in all, verb groups tend to be most frequent in the contexts of SIR. The patterns that are found in sentences with high number of in-text references ($N \geq 5$) are mostly composed of noun groups. Finally, we note that the majority of in-text references appear at the end of the sentences for both MIR and SIR. In our experiment, we considered both the left and right contexts around the in-text references. However, only the left contexts appear among the top 10 most frequent contexts. This implies that there are much fewer right contexts in the corpus due to the fact that the in-text references are often at final position in the sentence.

5 Discussion and conclusion

We have proposed a study of multiple in-text references (MIR) in respect of their positions in the rhetorical structure

[6] PLOS ONE, 2006, DOI: 10.1371/journal.pone.0000006.
[7] PLOS ONE, 2006, DOI: 10.1371/journal.pone.0000024.

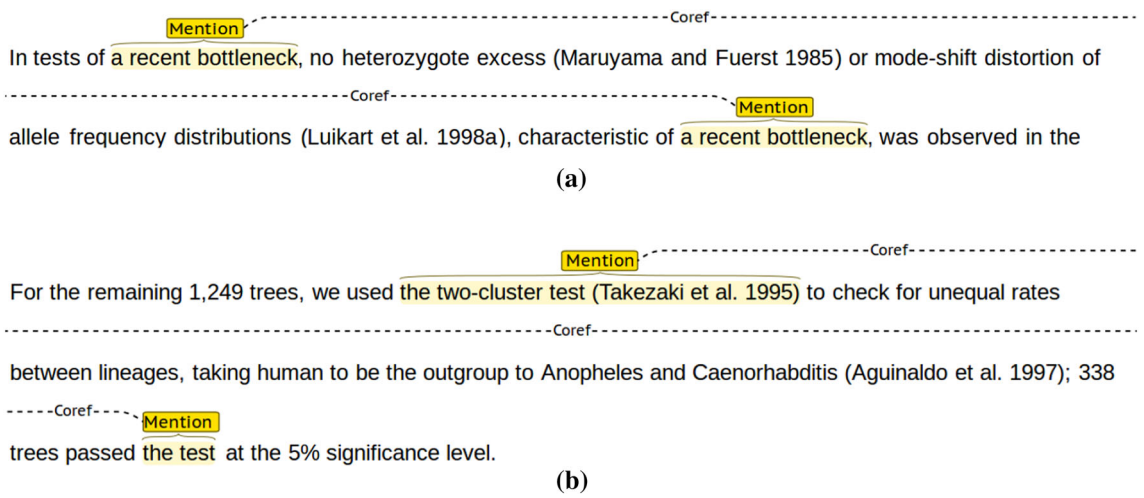

Fig. 4 Coreference resolution examples

of articles and their syntactic contexts. By studying a large-scale corpus of about 80,000 articles published by PLOS, we show the following key points:

- MIR are rather frequent in all sections of articles: 41% of the sentences with citations contain MIR.
- In the Introduction section, MIR account for more than half of the sentences containing citations.
- The MIR with two elements are the most homogeneous: they appear quite often in the Introduction, Results and Discussion sections (about 20% of sentences with citations) and in about 15% of sentences with citations in the Methods section.
- There exist sentences with very high number of in-text citations (more than 100). Such sentences are specific to the domain of medicine and the systematic review article type.
- SIR and MIR appear for the most part at the end of the sentences.
- In the Methods section, SIR and MIR are most often part of the verb group in the sentence, while in the other sections they are surrounded by noun groups.
- MIR are most likely to occur near verb groups in sentences containing few in-text references ($N \leq 5$) and in the Methods section.

In this study, the notion of MIR raises the question of the importance and the role of MIR in scientific articles. The implications of this study are relevant from the perspective of networks as bibliographic coupling [16], clustering [23] and co-citation [22], but also for the analysis of the functions of citations. Furthermore, many applications can benefit from the differentiation between single and multiple references such as automatic summarization [11,21] or automatic generation of surveys [20]. More generally, the distribution of MIR along the text progression has important implications for understanding the contexts of citations. For example, we can consider the following sentence:

> *Previous attempts to apply functional genomics methods to address these questions used various approaches, including* ***DNA microarrays*** *(Hayward et al. 2000; Ben Mamoun et al. 2001; Le Roch et al. 2002),* ***serial analysis of gene expression*** *(Patankar et al. 2001), and* ***mass spectrometry*** *(Florens et al. 2002; Lasonder et al. 2002) on a limited number of samples from different developmental stages.*[8]

In this sentence, there are 3 aggregates of in-text references and each aggregate is characterized by a noun group that identifies topics related to the references. The automatic identification of these topics will allow to assign them to each of the in-text references. This example shows that work at the level of sentences is not enough if we want to obtain fine and accurate results for content citation analysis.

The observations of this study suggest the presence of MIR implies the existence of features such as topics, keywords, methods that are common to all works cited in an aggregate of in-text references. This means that by examining the text content of such sentences one can obtain information on the topics that are shared by the group of cited works. The analysis of the syntactic patterns is a first step to solve this problem. In fact, we have observed that MIR appear most often as parts of noun groups. These noun groups, once identified and extracted, could be considered as candidates for topics related to the groups of cited works.

Furthermore, in our study we do not take into account the possible anaphoras that can be significant when determining the topics related to in-text references. This aspect is essen-

[8] PLOS Biology, 2003, DOI: 10.1371/journal.pbio.0000005.

tial in the assignment of characteristics, attributes, semantic relations or topics of research. For this reason, Coreference Resolution is relevant for the study of MIR, as it allows to consider the links between different textual units and their interactions. For example, Fig. 4 shows such links in sentences containing MIR.

The possibility to establish relations in the contexts of MIR and to identify the topics related to MIR is important for various applications around the exploitation of scientific corpora. Indeed, many works in automatic summarization extract information from in-text reference contexts in order to rebuild a summary of a cited paper. It is important also to analyze the relations between in-text references in the same sentence that can have different roles. High precision topic identification is crucial for information retrieval and recommender systems. For these reasons, the study of MIR and their contexts is essential for the good understanding of the significance of citations.

Acknowledgements We thank Benoit Macaluso of the Observatoire des Sciences et des Technologies (OST), Montreal, Canada, for harvesting and providing the PLOS dataset.

References

1. Atanassova, I., Bertin, M.: Temporal properties of recurring in-text references. D-lib Magazine 22(9/10) (September/October 2016)
2. Bertin, M., Atanassova, I.: A study of lexical distribution in citation contexts through the IMRaD standard. In: Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval (ECIR 2014), pp. 5–12, Amsterdam, 13 April 2014
3. Bertin, M., Atanassova, I., Larivière, V., Gingras, Y.: The invariant distribution of references in scientific papers. J. Assoc. Inf. Sci. Technol. **67**(1), 164–177 (2016)
4. Cronin, B.: The need for a theory of citing. J. Doc. **37**(1), 16–24 (1981)
5. Cronin, B.: Semiotics and evaluative bibliometrics. J. Doc. **56**(4), 440–453 (2000)
6. Cronin, B., Sugimoto, C.R.: Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. MIT Press, Cambridge (2014)
7. Day, R.: Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, chap. The Data—it is Me!. MIT Press, Cambridge (2014)
8. Demarest, B., Sugimoto, C.R.: Argue, observe, assess: measuring disciplinary identities and differences through socio-epistemic discourse. J. Assoc. Inf. Sci. Technol. **66**(7), 1374–1387 (2015)
9. Ding, Y., Liu, X., Guo, C., Cronin, B.: The distribution of references across texts: some implications for citation analysis. J. Informetr. **7**(3), 583–592 (2013)
10. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., Zhai, C.: Content-based citation analysis: the next generation of citation analysis. J. Assoc. Inf. Sci. Technol. **65**(9), 1820–1833 (2014)
11. Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., Radev, D.: Blind men and elephants: what do citation summaries tell us about a research article? J. Am. Soc. Inf. Sci. Technol. **59**(1), 51–62 (2008)
12. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 363–370. Association for Computational Linguistics (2005)
13. Gipp, B., Beel, J.: Citation Proximity Analysis (CPA) a new approach for identifying related work based on co-citation analysis. In: Larsen B., Leta J. (eds.) 12th International Conference on Scientometrics and Informetrics, vol. 2, pp. 571–575. International Society for Scientometrics and Informetrics, Rio de Janeiro, 14–17 July 2009
14. Hu, Z., Chen, C., Liu, Z.: The recurrence of citations within a scientific article. In: Salah A., Tonta A., Akdag Salah C., Sugimoto U.A. (eds.) 15th International Society of Scientometrics and Informetrics Conference. International Society for Scientometrics and Infometrics, Bogazii University Printhouse, Istanbul, June 29–July 3 2015
15. Kaplan, D., Tokunaga, T., Teufel, S.: Citation block determination using textual coherence. J. Inf. Process. **24**(3), 540–553 (2016)
16. Kessler, M.M.: Bibliographic coupling between scientific papers. Am. Doc. **14**(1), 10–25 (1963)
17. Liu, S., Chen, C.: The proximity of co-citation. Scientometrics **91**(2), 495–511 (2011)
18. Liu, S., Chen, C.: The effects of co-citation proximity on co-citation analysis. In: 13th Conference of the International Society for Scientometrics and Informetrics, vol. 1 and 2, pp. 474–484. International Society for Scientometrics and Infometrics, Durban, 4–7 July 2011
19. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: ACL (System Demonstrations), pp. 55–60 (2014)
20. Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., Zajic, D.: Using citations to generate surveys of scientific paradigms. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 584–592. Association for Computational Linguistics, Boulder, May 31–June 5 2009
21. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22 International Conference on Computational Linguistics, vol. 1, pp. 689–696. Association for Computational Linguistics, Manchester, 18–22 Aug 2008
22. Small, H.: Co-citation in the scientific literature: a new measure of the relationship between two documents. J. Am. Soc. Inf. Sci. **24**(4), 265–269 (1973)
23. Small, H., Sweeney, E., Greenlee, E.: Clustering the science citation index using co-citations. II. Mapping science. Scientometrics **8**(5–6), 321–340 (1985)
24. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics (2012)
25. Wouters, P.F.: The citation culture. Ph.D. thesis, University of Amsterdam (1999)
26. Zhao, D., Strotmann, A.: Re-citation analysis: Promising for research evaluation, knowledge network analysis, knowledge representation and information retrieval? In: 15th International Society of Scientometrics and Informatics Conference, pp. 1061–1065. International Society for Scientometrics and Infometrics, Bogazii University Printhouse, Istanbul, June 29–July 3 2015